ARMY RESEARCH LABORATORY

Divergence Measures Tool: An Introduction with Brief Tutorial

by Douglas M. Briesch, Claire E. Jaja, Terrence J. Moore, and Clare R. Voss

ARL-TN-0599 **March 2014**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TN-0599 **March 2014**

Divergence Measures Tool: An Introduction with Brief Tutorial

Douglas M. Briesch, Claire E. Jaja, Terrence J. Moore, and Clare R. Voss
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
March 2014	Final	October 2012 to August 2013

4. TITLE AND SUBTITLE
Divergence Measures Tool: An Introduction with Brief Tutorial

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Douglas M. Briesch, Claire E. Jaja, Terrence J. Moore, and Clare R. Voss

5d. PROJECT NUMBER
R.0006155.19

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-T
2800 Powder Mill Road
Adelphi MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-0599

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report provides new users of the Divergence Measures Tool (DMTool) with an overview of its core functionality. The DMTool calibrates natural language text corpora for "how dissimilar" they are from each other, based on the distribution of the relative frequencies of the terms in each corpus. Computation involves a suite of seven information-theoretic divergence measures calculated on given pairs of text files. Users are provided with the resulting scores and list views of the file terms with their frequencies, as used in computing the scores. Use cases and a hands-on tutorial are provided in this report.

15. SUBJECT TERMS
Divergence measures, natural language processing, text corpus, text domain, text genre, term frequency

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Douglas M. Briesch
Unclassified	Unclassified	Unclassified	UU	42	19b. TELEPHONE NUMBER (*Include area code*) (301) 394-1057

Contents

List of Figures

INTENTIONALLY LEFT BLANK.

1. Introduction

This report provides new users of the Divergence Measures Tool (DMTool) with a brief overview of its core functionality and a hands-on tutorial with two files to work with the tool immediately. The DMTool at this stage is a work-in-progress: it supports our in-house U.S. Army Research Laboratory (ARL) research, as initially reported in Jaja et al. (2012).[*] An extended DMTool report that covers the implementation of the measurements and recent extensions is forthcoming.[†]

The initial impetus for building this software application was our work with natural language (NL) text classifiers. As we began examining available text datasets for training and testing the two-way classifiers that we wanted to build, it became apparent that we would need to construct additional text corpora and calibrate them for "how dissimilar" they were from each other.[‡] Since we did not know a priori which measures would be adequately sensitive in detecting differences across a wide range of Arabic-language text files (they varied by genre, domain, spelling variation, size, etc.), our initial requirement was for a tool that would support this calibration evaluation task.

As a result, the DMTool runs a suite of seven information-theoretic divergence measures on given pairs of user files (called P and Q) and provides users with both the resulting scores and five list views of the file terms with their frequencies, as used in computing the scores. Each list view can toggle between an alphabetic and frequency-based ordering of all terms in that view. The five views show all terms and their frequencies: in the P file, in the Q file, in only the P file and not the Q file, in only the Q file and not the P file, and in both the P and Q files.

The measures all compare—each in a slightly different way—the distribution of the relative frequency of the terms in the files.[§] The basic capability of the DMTool includes seven divergence measures: Rényi, Kullback-Leibler, Bhattacharyya, Jensen-Shannon, Variational (also known as L1, taxicab, Manhattan, city-block), Euclidean (also known as L2), and Cosine (Rényi, 1961; Kullback and Leibler, 1951; Bhattacharyya, 1943; Singhal, 2001; Lin, 1991; Huang, 2008). For mathematical definitions of the measures, see appendix A.

[*] We welcome all feedback on the tool so that we can continue to improve it.

[†] The Batch tab and Super-batch tab overviewed in section 7 have been extended and refined over the last year as part of our ongoing collaboration with Dr. Terrence Moore (ARL).

[‡] We followed the standard assumption in the field of computational linguistics that the more dissimilar the files, the better the classifiers would perform in correctly categorizing the files.

[§] These measures are string-based: they do *not* provide any deeper lexical, semantic, or conceptual analyses of the terms. They simply count the terms and have no basis for understanding the meaning(s) of these terms, or their relation to each other.

These measures have been put to many uses in natural language processing (NLP). In the evaluation of machine translation (MT) engines, for example, researchers have begun to estimate the quality of the MT output for a previously unseen input text by calculating the divergence between the MT training data and the new input text. For the scenario where multiple MT engines trained on different data are available at MT runtime, these measures may help determine which engine will provide the best results for a new text set.

The tool is organized by tabs. The user can always see all tabs by name, like a traditional filing system, where the names appear individually lined up across the top of the interface. Only one tab is selected and open at a time, with its full screen and functionality available to the user. To change the view from one tab to another, the user can simply click on a different tab to open that one and close the current one. (There is no automated workflow between tabs.)

The **Basic** tab allows the user to upload, preprocess the text, visually compare the terms in the two *corpora,* and calculate the divergence measures on the two corpora.

The **Help** tab provides ready access to information on the underlying equations and definitions used in the tool.

Several other tabs provide support for other use cases. We include in this introduction an overview of the **Classifier** tab that was the original use case motivating the construction of this tool. The functionality of the other tabs in the DMTool are described briefly in section 7 of this report and will be covered in more detail in the forthcoming extended DMTool technical report. Specific user questions are addressed in appendix D of this report.

2. User Manual Conventions

In this report, we adhere to the following typographical conventions to distinguish among different types of information:

- Words with technical definitions (as spelled out in section 5) are italicized.
- References to sub-sections in this user manual are bolded and italicized. In Microsoft Word, all internal section references also work as links to the referenced section or sub-section when clicked while holding down the Ctrl button.
- References to labels from the tool are bolded and capitalized.
- References to input or output for the tool are displayed in Courier New typeface.

To allow for ongoing documentation of the DMTool over time as it has evolved, we made the following decisions in creating the screenshots:

- Many screenshots are cropped down to that portion of screen layout most relevant to the accompanying description. Cropped screenshots in figures in sections 6, 7, and 8 of the manual, correspond respectively to partial views of the **Basic**, **Classifier**, or **Help** tab screens.
- Later versions of the DMTool have more tabs. The use cases that these tabs support are described in section 7, with details for running these tabs in the extended technical report.

3. Requirements

The following are the requirements for the tool:

- This tool is compatible with Windows XP, Windows Vista, and Windows 7.
- Any files uploaded into the tool must be .txt files in ASCII or UTF-8 format.
- This tool has been tested on English and Arabic script[**], but should work on any other language with ASCII or UTF-8 encoding.
- For best results, the input text should have its punctuation tokenized, i.e., separating each punctuation mark from a word it is attached to, by adding in an extra blank space.[††] The input text should be lowercased, when applicable. Both of these preprocessing options are available within the tool itself when the user uploads text (see figure 2).
- The divergence measure calculations are intended to distinguish the distribution of words in two sets. The smaller the sets being compared, the less likely they are to capture the true distributions from which they were drawn. **We caution the user here and leave it to each individual to determine the appropriate set sizes for their application**.

4. Installation and Setup

Included with this tool should be four files:

- DMT.exe
- Sample 1.txt
- Sample 2.txt

[**] So far, this has included Modern Standard Arabic, Farsi, Dari, Urdu, and Pashto.

[††] Exceptions to this approach may include preserving numeric expressions.

- User Manual.doc

Place these files together in the folder where the tool is stored. To run the tool, simply double click the "DMT.exe" file. A tutorial on Sample 1.txt and Sample 2.txt is provided in appendix C.

5. Definitions

The following are definitions used in this manual in describing the tool:

- *type*: unique word in a given text
- *token*: any instance of a word in a given text
- *corpus* (pl. *corpora*): a collection of written texts

6. Basic Tab

The **Basic** tab is the default tab that opens when the tool is started. It provides for the following operations.

6.1 Uploading User Files

At the top left of the screen, as shown in figure 1, is where one may upload two text files with the option to name them each with a shorter label. To find and upload the files from a computer, click the "…" button to the right of the **File** fields and browse to the desired files. If short labels are entered into the **Name** fields, these names will appear (in lieu of the source file names) in the DMTool's output reports (see section 6.7 ***Reports***). The names cannot be identical, must be 27 or fewer characters, and cannot include the following symbols: [ ] * ? / \ . The output folder to hold any reports generated by the DMTool will be located, by default, in the user's Documents folder. This may be changed if desired, by clicking the "…" button to the right of the **Output Folder** field, browsing until locating another folder to use instead, and then selecting that desired folder's name.

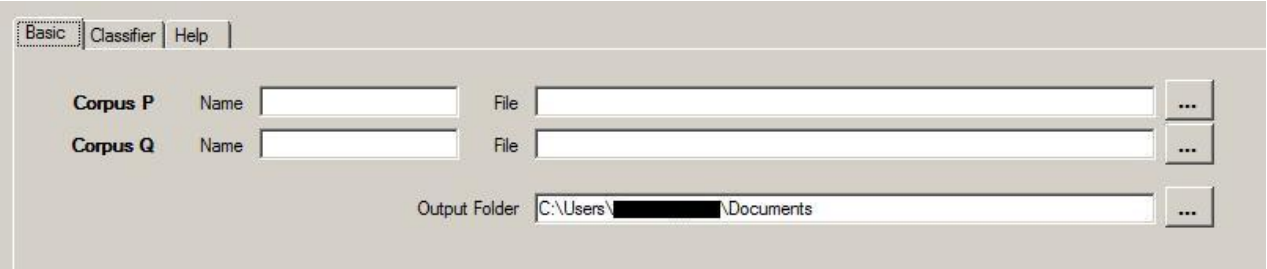


Figure 1. Uploading files with optional short names, and designating path for output files (**Basic** tab).

6.2 Preprocessing Text in User Files

To the right of the **Basic** tab screen (see also section 6.1 ***Uploading User Files***) are the **Preprocessing** options, as shown in figure 2. By default, these are unselected. There are two options available:

- **Tokenize Punctuation** will separate any word-external punctuation from the word to which it is attached; this ensures that a word directly before and/or directly after a comma, period, quotation mark, or other punctuation will be recognized after pre-preprocessing and added to the count for that word elsewhere, and the punctuation itself will be counted separately. By applying this process to word-external punctuation only, this avoids the incorrect separation of other punctuation, as would otherwise happen with "can't" becoming three *tokens*: `can ' t.` (For a full list of what is recognized as punctuation by this tool, see appendix B.)
- **Convert to Lowercase** will convert all characters in Roman script to lowercase; this ensures that `The` and `the` are counted as the same word. (The user must, however, determine when this is not applicable, such as when the Roman script is the Buckwalter transliteration of Arabic script.)

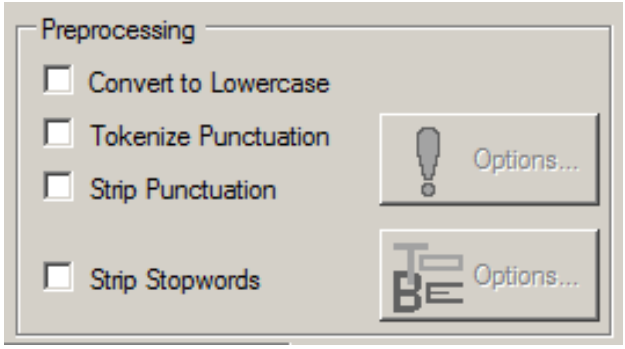


Figure 2. Altering input text before calculating measures (**Basic** tab).

6.3 Display Options for List Views

At the right of the **Basic** tab screen, below the **Compare Corpora** button (see section 6.4 ***Compare Corpora Button***), there are three checkboxes with options for text terms displayed on the screen within the tool (as shown in figure 3):

- **Right to Left** switches the text direction for languages such as Arabic; this option is unselected by default.
- **Use Buckwalter decoding** takes Arabic that has been written in Roman script using Tim Buckwalter's transliteration schema (Buckwalter, 2002) and converts it into Arabic script[‡‡]; this option is also unselected by default.
- **Automatically Sort by Frequency** sorts the word lists by the *type* frequencies; this option is selected by default. However, the user may wish to deselect this option if the files are especially large and the user is mainly concerned with the divergence measures because the word list sorting can be memory intensive and slow down the processing.

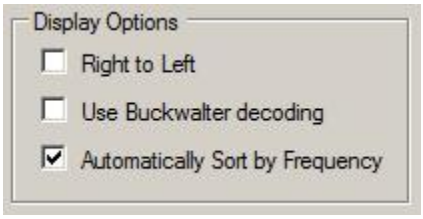


Figure 3. Screen display options in list views (**Basic** tab).

6.4 Compare Corpora Button

After the user has loaded the files and adjusted any preprocessing or display options desired, click on the **Compare Corpora** button (figure 4) to run the back-end calculations and generate the word list views and divergence measures (see section 6.5 ***Word Lists*** and section 6.6 ***Divergence Measures***).

Figure 4. Button to run divergence measures calculations (**Basic** tab).

[‡‡]This step performs a character-for-character substitution.

6.5 Word Lists

After loading files, the three blocks, with different color backgrounds as shown in figure 5, labeled **Corpus P**, **Corpus Q**, and **Intersection** will fill with words from the uploaded text. Within the top blocks for **Corpus P** and **Corpus Q,** there are two panels, or list views — one view with all of the words in the corpus and their frequencies, another view with the words that occur only in one corpus but not the other. As indicated by the name, the **Intersection** block below lists only words that occur in both corpora.

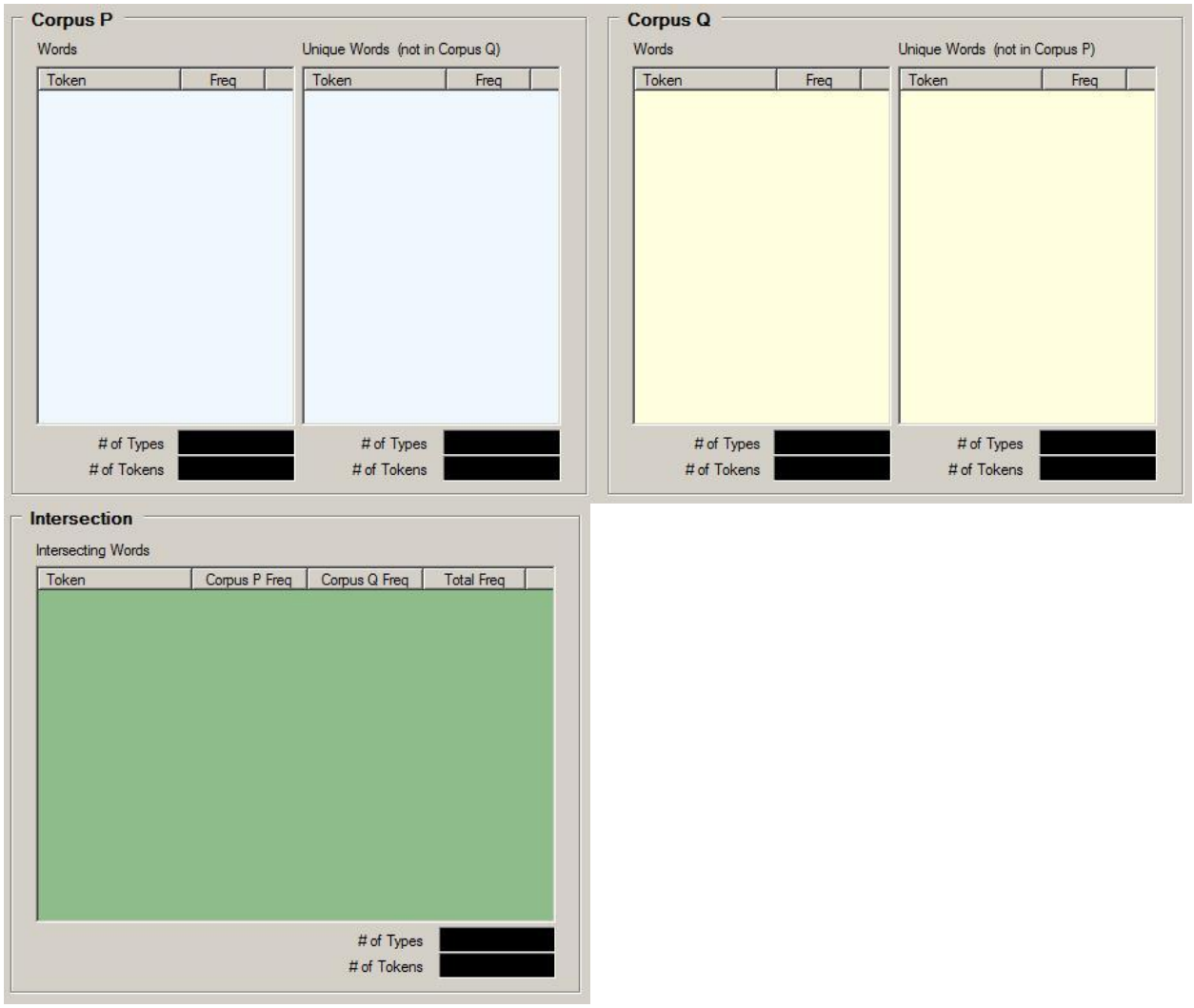


Figure 5. Results in different list views (**Basic** tab).

At the bottom of each list view are two summary counts for that set: the number of types in the set and the number of tokens in the set. The number of types corresponds to the number of rows in the list view above. The number of tokens corresponds to the sum of values in the frequency column in the list view above.

6.6 Divergence Measures

The bottom right box labeled **Divergence Measures**, as shown in figure 6, provides the nine divergence measure calculations (the Rényi and Kullback-Leibler measures are asymmetric and are thus calculated in both directions). The Rényi measure uses a constant Alpha; this is set to 0.99 by default but can be any number greater than 0 except for 1. All of the measures have been normalized as difference measures[§§] that return 0 when the two *corpora* are identical. The range of each measure is listed on the right side to help the user interpret the numbers.

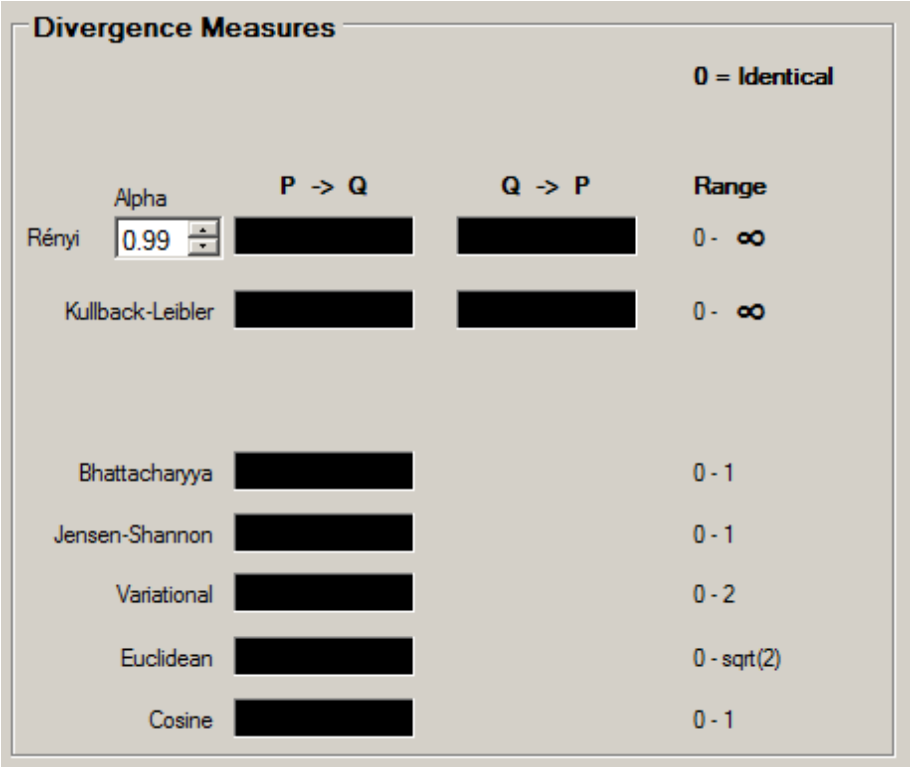


Figure 6. Resulting measures (**Batch** tab).

6.7 Reports

On the far right side of the screen is a box labeled **Reports** (figure 7). The button in this box generates Excel spreadsheets as follows:

- Selecting the **Word Distribution** checkbox and then clicking the **Generate Report** button will create a spreadsheet with the information from the word lists (see section 6.5 ***Word Lists***).

[§§] This applies for the cases where the measures have been traditionally defined as similarity measures and the value of 1 indicates identity of the two sets.

- Selecting the **Divergence Measures** checkbox and then clicking the **Generate Report** button will create a spreadsheet with the calculated divergence measures (see section 6.6 ***Divergence Measures***).
- Selecting both checkboxes will generate both spreadsheets.

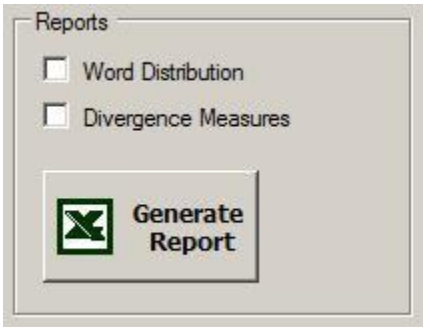


Figure 7. Generating results in summary reports (**Basic** tab).

6.8 Clear Data Button

The **Clear Data** button (figure 8) located in the lower right corner of the screen serves to clear all of the previously loaded data. This button ensures that no data from previous files is still stored in memory.

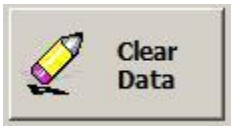


Figure 8. Button to remove all values in all fields on the screen and in memory (**Basic** tab).

7. Use Cases

As the DMTool has evolved, new tabs have been added to the interface to support new use cases. In this section, we describe the **Classifier** tab in detail and then provide a brief overview of the other use cases for which the DMTool has been augmented. Details for running these latter cases with the **Batch** and **Super Batch** tabs are not necessary for getting started using the DMTool and so have not been included in this report, but will appear in the extended technical report.

7.1 Classifier Construction and Evaluation

The construction of a two-way classifier requires at a minimum four files: two as training sets representative of the two domains during the classifier development stage and two as test sets again as separate representatives of the two domains during the classifier evaluation stage. The

Classifier tab supports all pair-wise comparisons of these files so that the user can determine how to select and partition the available corpora for training and testing.

The **Classifier** screen is accessed by selecting the tab at the top labeled **Classifier**. This screen also supports the visualization of divergence scores within a corpus. Most of this screen contains the same controls, buttons, and displays as described above for the **Basic** screen in section 6 (specifically, ***Preprocessing***, ***Display Options***, ***Compare Corpora Button***, ***Word Lists***, ***Divergence Measures***, and ***Clear Data Button***).

At the top of the **Classifier** screen (figure 9), the user uploads four files from two different domains (Domain 0 and Domain 1). For each domain, there is one training file and one test file. The domain names cannot be identical, must be 11 or fewer characters, and cannot include the following symbols: [ ] * ? / \ . As with the **Basic** screen, the DMTool provides a default path and name in the **Output Folder** field where the user can either directly enter or browse and select the folder name to where output results will be stored.

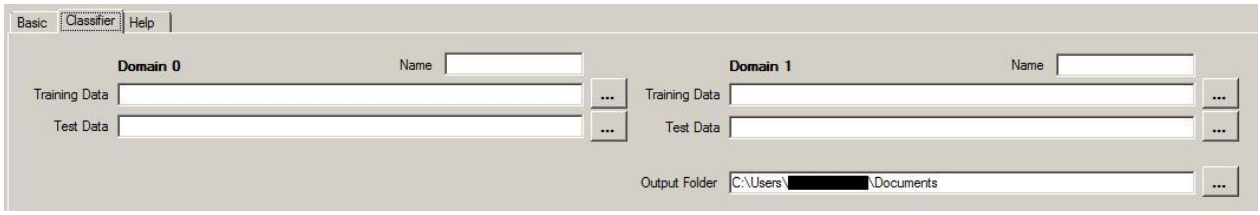


Figure 9. Uploading files (**Classifier** tab).

7.1.1 Corpus P and Corpus Q Selection

After all four files are loaded, the user needs to select the two to be compared as Corpus P and Corpus Q, as is done in the **Basic** tab.

To do so, there is a drop-down menu for each corpus just above the word lists, as shown in figure 10. These menus are automatically populated after the load, so all the user needs to do is select which two files to compare at any given time. The word list views, as well as the measure calculations, will change depending on the selected files.

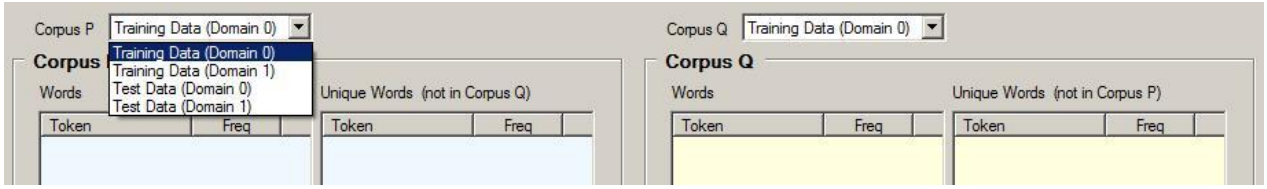


Figure 10. Selecting the files to be compared (**Classifier** tab).

7.1.2 Reports

The Reports portion of the **Classifier** screen differs by one option from what has already been introduced for generating reports on the **Basic** screen: there is a drop-down menu linked to the Divergence Measures checkbox. The user can use this menu to generate a report (figure 11) for either the current pair of files (selected by the drop-down menus described in section 7.1.2 ***Corpus P and Corpus Q Selection***) or for all six possible combinations of different files (Training Domain 0 – Training Domain 1, Test Domain 0 – Test Domain 1, Training Domain 0 – Test Domain 0, Training Domain 1 – Test Domain 1, Training Domain 0 – Test Domain 1, Training Domain 1 – Test Domain 0).

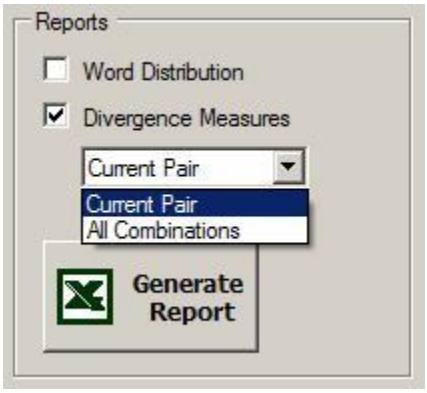


Figure 11. Generating results in summary reports (**Classifier** tab).

7.2 Within Corpus Analysis

In the process of constructing MT engines that would translate a set of military manuals from Arabic to English, we created numerous subsamples from the manuals and discovered that the resulting MT engines yielded a wider range of evaluation scores than we had been expecting. This led us to ask how widely different portions of the manuals varied from each other. After formulating the question this way, we realized that we could partition the English side of the corpus into separate files and then—by extending the DMTool with a **Batch** tab to run multiple pair-wise comparisons—we could score each military manual partition against each of the other partitions and inspect how much their lexical content varied.

7.2.1 Uploading Files

The **Batch** tab was added to the later extended versions of the DMTool to support both this within-corpus analysis and the more general NxM dataset comparisons. As shown in figure 12, rather than using the **Basic** tab, the user can select the **Batch** tab and directly upload one collection of their files under the P Corpus Set (where P contains N files) and the other collection of their other files under the Q Corpus Set (where Q contains M files, possible the same as in P if running a within corpus analysis).

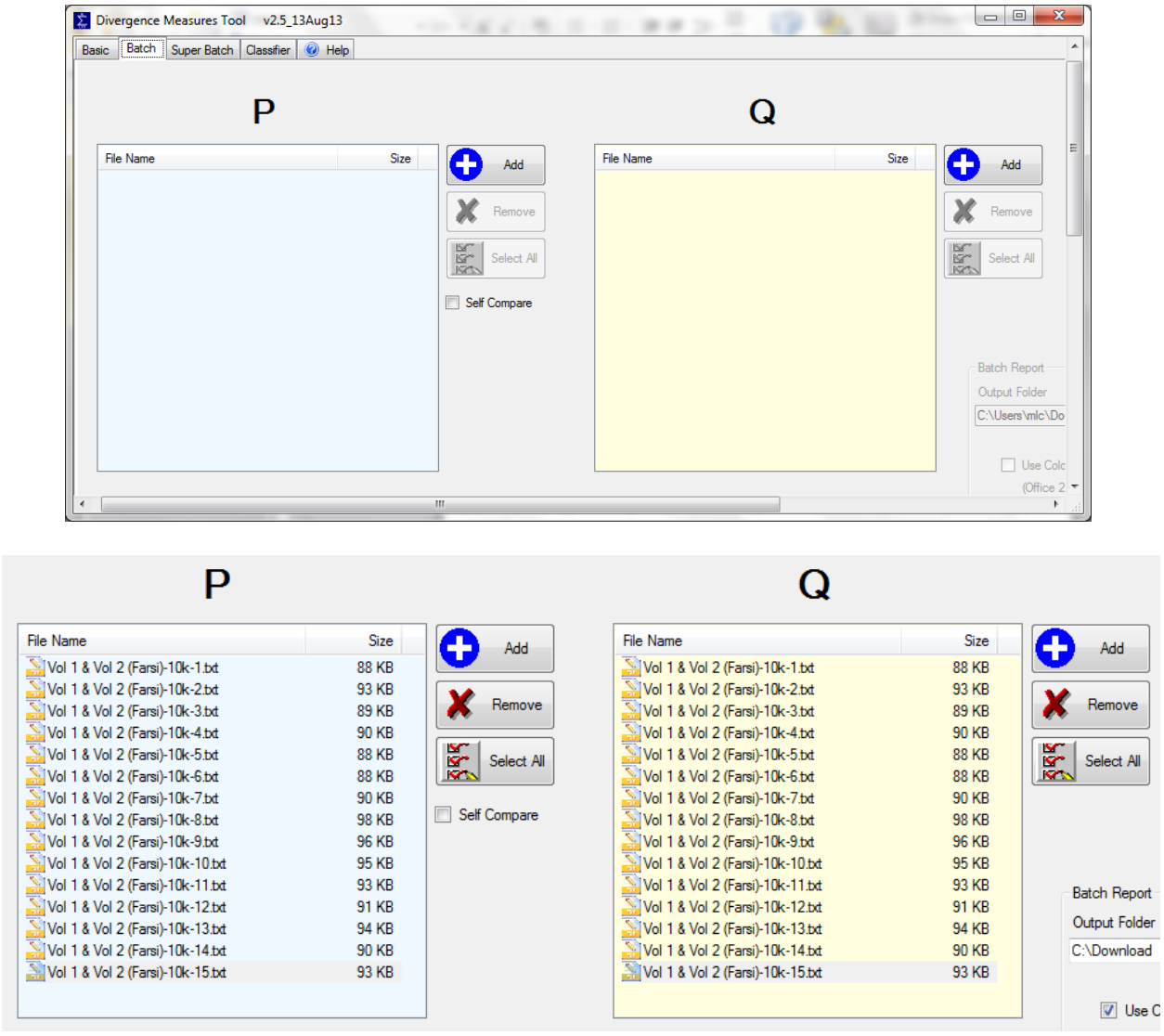


Figure 12. Uploading files: top screen before upload, bottom screen after a partitioned upload (**Batch** tab).

Instead of selecting one file for corpus P and one file for corpus Q, the user can select multiple corpora for P and for Q by using the **Add** button. The **Remove** button will remove all of the files that are selected. The **Select All** button will select all of the files listed in the list box. When the user clicks on the **Compare Corpora** button, the program will compare all of the pairwise combinations possible between the list of corpora in P and the list of corpora in Q.

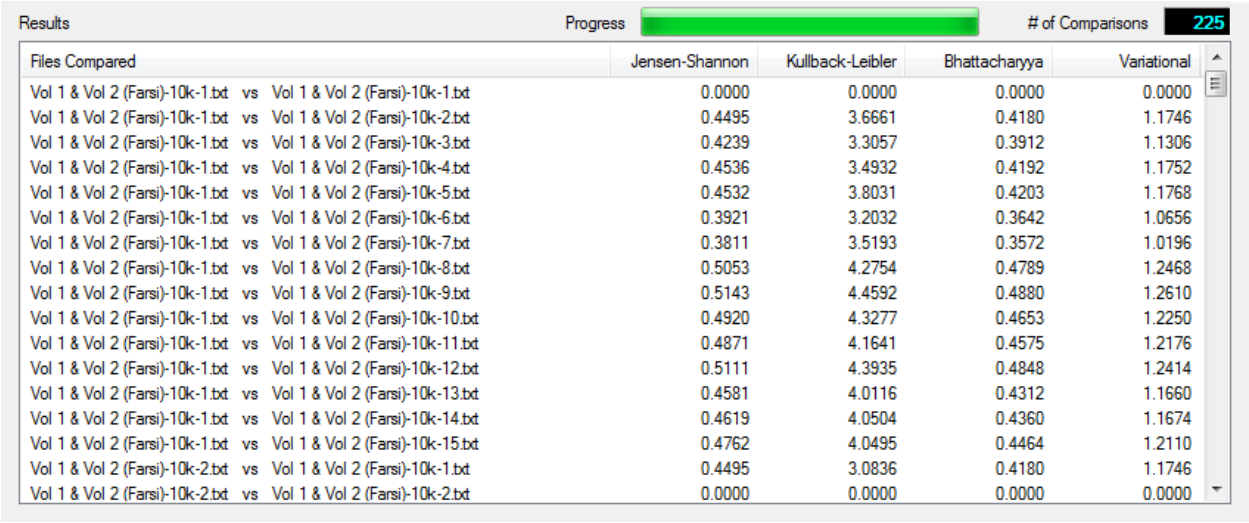
Another variation on a within-corpus analysis is provided by the **Self Compare** checkbox. This functionality enables the user to perform a unique "hold-one-out" type of comparison. These comparisons are run between each individual file in corpus P (each as a "hold out") against one

new file that is constructed "on-the-fly" by combining **all of the other** ("non-held out") files in corpus P.

7.2.2 Results

The **Batch** tab is very similar in functionality to the **Basic** tab but with only four measures: Jensen-Shannon, Kullback-Leibler, Bhattacharyya, and Variational. Due to screen space constraints, no individual word breakdowns are displayed. If desired, these can, of course, be individually computed in the **Basic** tab.

Figure 13 shows the screen display with results for the four measures calculated on each pairwise file comparison, following the upload in figure 12. (Notice, for example, that partitions 1 and 9 are the most divergent pair, with the highest scores in the ninth row within each column.)



Files Compared	Jensen-Shannon	Kullback-Leibler	Bhattacharyya	Variational
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-1.txt	0.0000	0.0000	0.0000	0.0000
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-2.txt	0.4495	3.6661	0.4180	1.1746
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-3.txt	0.4239	3.3057	0.3912	1.1306
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-4.txt	0.4536	3.4932	0.4192	1.1752
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-5.txt	0.4532	3.8031	0.4203	1.1768
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-6.txt	0.3921	3.2032	0.3642	1.0656
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-7.txt	0.3811	3.5193	0.3572	1.0196
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-8.txt	0.5053	4.2754	0.4789	1.2468
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-9.txt	0.5143	4.4592	0.4880	1.2610
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-10.txt	0.4920	4.3277	0.4653	1.2250
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-11.txt	0.4871	4.1641	0.4575	1.2176
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-12.txt	0.5111	4.3935	0.4848	1.2414
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-13.txt	0.4581	4.0116	0.4312	1.1660
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-14.txt	0.4619	4.0504	0.4360	1.1674
Vol 1 & Vol 2 (Farsi)-10k-1.txt vs Vol 1 & Vol 2 (Farsi)-10k-15.txt	0.4762	4.0495	0.4464	1.2110
Vol 1 & Vol 2 (Farsi)-10k-2.txt vs Vol 1 & Vol 2 (Farsi)-10k-1.txt	0.4495	3.0836	0.4180	1.1746
Vol 1 & Vol 2 (Farsi)-10k-2.txt vs Vol 1 & Vol 2 (Farsi)-10k-2.txt	0.0000	0.0000	0.0000	0.0000

Figure 13. Results of file comparisons (**Batch** tab).

7.2.3 Measures

The **Measures** area, as shown in figure 14, will display the Mean, Minimum value, Maximum value, and the Standard Deviation for each of the four measures. These values are calculated over the whole list of comparisons shown in the **Results** window.

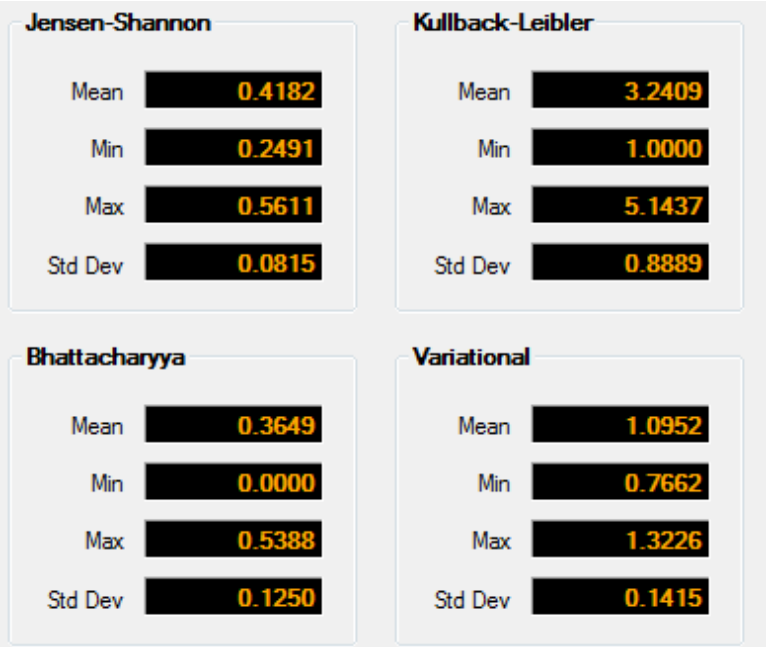


Figure 14. Summary results (**Batch** tab).

7.2.4 Batch Report

The **Batch Report** allows the user to select an output directory and then generate a report (figure 15). The report is an Excel spreadsheet with the first tab containing all data (as shown in the **Results** window and the **Measures** area) and the subsequent tabs in the spreadsheet containing a file-by-file listing of each of the measures' scores.

Another more accessible visualization of a within-corpus analysis is also available via the **Batch** tab. The user can select the **Use Color Scales** checkbox feature to generate an automated color shading of the file-by-file comparison scores in a matrix, as shown in figure 16, ranging from the lowest scores (most similar or least divergent comparisons) as greenest to the highest scores (most dissimilar) as the reddest. This feature is only available with Microsoft Office version 2007 and above.

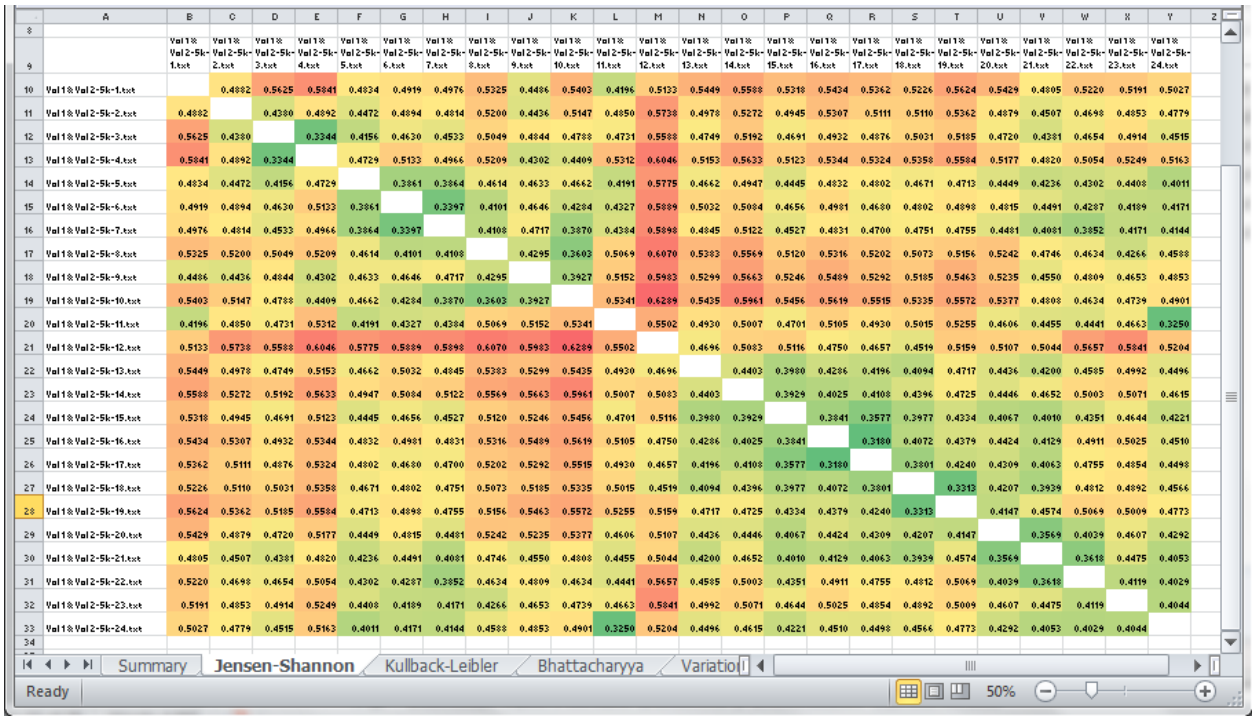


Figure 15. Generating summary results (**Batch Report**).

Figure 16. Summary of measures in color-coded matrix (**Batch** tab).

The colors in this figure, for example, suggest that the first 11 partitions and the last 12 partitions are more distinct from each other than they are alike. The ground truth for this corpus corroborates this interpretation: these two sets correspond to two volumes of a medical textbook.

Furthermore, the green cell that appears to be an island in the matrix in the comparison of the last partition of the first set (filename ends -11) and the last partition of the second set (filename ends -24) is also quite revealing: there is an index of terms at the end of the first textbook volume that contains many terms in the index at the end of the second textbook volume.

7.3 Assessment of Line-Aligned Parallel Files

Not long after extending the DMTool for the within-corpus analysis in building MT engines, we also realized that the DMTool could help us detect potential errors in the line-level alignments of

parallel source-to-target language files that are the training and test datasets for MT efforts. We leave it as an exercise for the eager reader to discover how the **Super Batch** tab, as described in this section, supports this use case.

The **Super Batch** tab allows the user first to select two corpora and have the DMTool automatically generate corpora partitions, either in terms of the total word count for each partition or the total number of bins for partitioning each corpus, and then to compare all of the pairwise combinations of those partitions. As with the **Batch** tab, the DMTool calculates the same four divergence measures for the **Super Batch** tab: Jensen-Shannon, Kullback-Leibler, Kullback-Leibler (reversed), and Variational.

7.3.1 Uploading Files

The **Uploading Files** block (figure 17) is where the user enters the paths for Corpus P and Corpus Q and the output folder, assigning short names as well if desired. The **Output Folder** will contain all of the Corpus P and Corpus Q partition files that are created, as well as any reports generated.

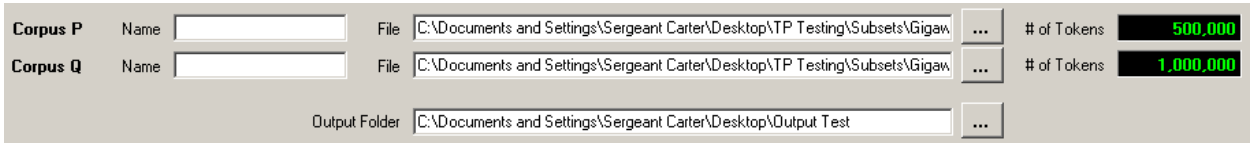


Figure 17. Uploading files (**Super Batch** tab).

7.3.2 Construction of Subsets

The **Subsets** block (figure 18) is where the user selects the size of the subsets to be created before comparison scores are calculated. If a corpus is not evenly divisible by the subset size selected for the partitions, the last subset created at the end of the corpus will have fewer words than the size selected. If the selected subset size is greater than the size of the full corpus, no subsets will be created. Alternatively, the user can select the number of bins to partition the corpus so that both corpora will have the same total number of partitions. These two methods of subset selection are mutually exclusive. The user can do one or the other, but not both.

Figure 18. Construction set size for subsets (**Super Batch** tab).

7.3.3 Status

The **Status** window (figure 19) displays the progress of the backend code performing the divergence calculations as it is executing in real time. Status messages provide updates as individual subsets are created and file comparisons are calculated. The completion of the **Super Batch** process generates a status message as well as an audible signal to alert the user that the process has finished (assuming, of course, that the sound on the user's computer is not muted).

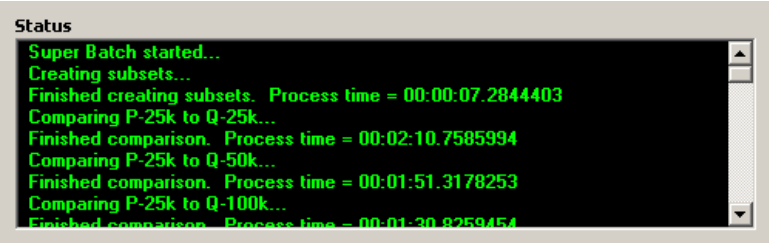


Figure 19. Status log window (**Super Batch** tab).

7.3.4 Scores

The **Scores** block is where the individual divergence measure scores are displayed for each pairwise comparison for four of the divergence measures: Jensen-Shannon, Kullback-Leibler, Bhattacharyya, and Variational (figure 20).

The **Data View** drop-down menu box, located on the upper right of the **Scores** block, allows the user to choose the score types to display, with the options being Mean, Median, Min-Max, Standard Deviation, and # of Comparisons. The **Mean** option shows the average of all pairwise comparison scores in each cell for the given partition size pairs. The **Median** option shows the value at the middle of the distribution of the scores in each cell. If the number of comparison scores is even, the middle two scores are averaged. The **Min-Max** option shows the lowest and highest divergence scores in each cell for the given partition size pairs. The **Standard Deviation** computes this value in each cell again for the given partition size pairs. The **# of Comparisons** option shows the number of pairs in each cell. The user also has a selection box, seen in blue, for each divergence measure, so that all individual pairwise scores computed for a particular cell may be displayed separately as well in the **Cell Details** block, as described in the next subsection of this report.

Figure 20. Scoring summary results (**Super Batch** tab).

7.3.5 Score Cells in Detail

Given that the summary result table collapses numerous individual comparisons, the **Cell Details** block enables the user to drill down and examine the individual divergence measure scores that went into the summary statistics for each of the selected subset comparison cells in the **Scores** block. An example of the detailed breakout is shown in figure 21 for the summary cells of P subsets of size 25K tokens scored against the Q subsets of size 25K tokens (for cells colored in blue in figure 20).

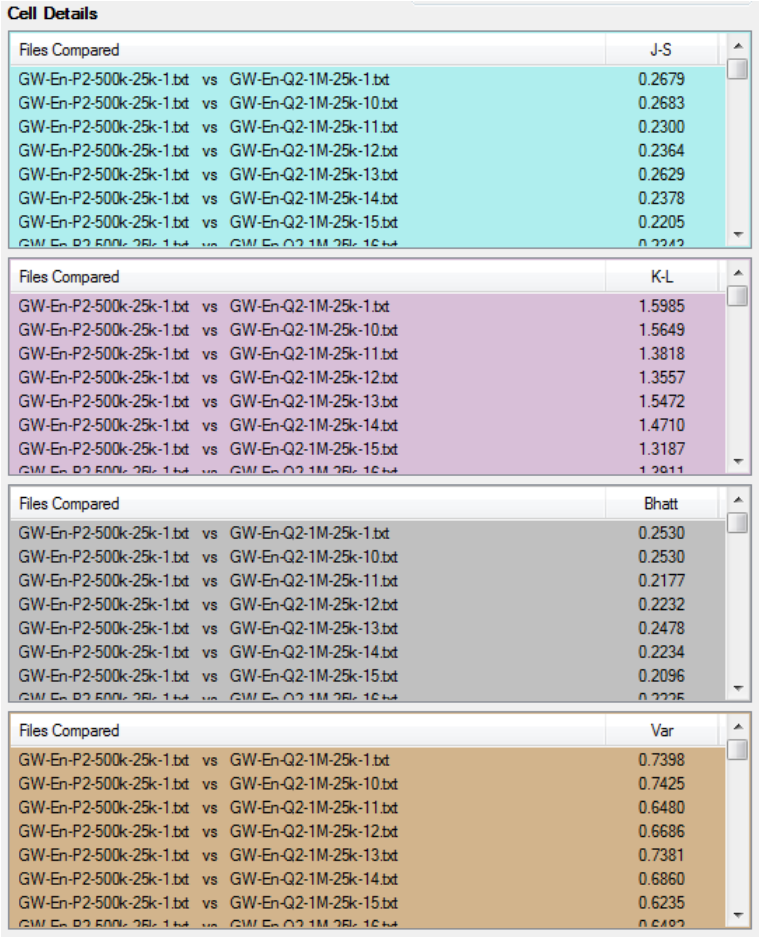


Figure 21. Cell-level details for pair-wise scoring results (**Super Batch** tab).

7.3.6 Reports

The **Reports** block of the **Super Batch** screen, as with the other screens, allows the user to select the type of report to be created and then to generate that report with the **Generate Report** button (figure 22). The two types of reports are both generated in Excel spreadsheets. The **Summary** report shows exactly what is displayed in the **Scores** block. The **Cell Details** report will show what is displayed in the **Cell Details** block, but only for those comparison scores from the subset comparison (cell) selected in the **Scores** block. If both report types are selected, only one Excel spreadsheet will be created, with the **Summary** data on the first tab and the **Cell Details** on the second tab of the same saved spreadsheet.

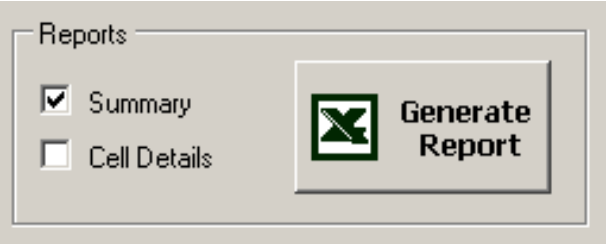


Figure 22. Generating reports (**Super Batch** tab).

8. Help Tab

The **Help** tab contains basic information on the tool, including the following two screen-internal tabs.

8.1 Assumptions and Definitions

This is the information on the assumptions and definitions as used in implementing the equations computationally. This includes the information from section 5 and appendix A.

8.2 Equations

These are the underlying mathematical equations for the measures. See appendix A.

9. Conclusion

The current version of the DMTool allows for basic functionalities for comparing text corpora on similarity measures. The following functionalities are under consideration for inclusion in future versions of this tool:

- Reading in files with UTF-16 encoding
- Automatically generating a Venn diagram comparing type counts of the compared files.
- Calculating the perplexity of each corpus.
- Eliminating current error when opening reports with Microsoft Excel 2007.
- Inputting a personalized punctuation list for the purpose of tokenization.
- Calculating the measures and word lists by n-grams.

10. References

Bhattacharyya, A. On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* **1943**, *35*, 99–109.

Buckwalter, T. Buckwalter Arabic Morphological Analyzer. Linguistic Data Consortium. (LDC2002L49), 2002.

Huang, A. Similarity Measures for Text Document Clustering. *New Zealand Computer Science Research Student Conference*, 2008.

Jaja, C.; Briesch, D.; Laoudi, J.; Voss, C. Assessing Divergence Measures for Automated Document Routing in an Adaptive MT System, In *Proceedings of the Language Resources and Evaluation Conference (LREC2012)*, Istanbul, Turkey, 2012.

Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Annals of Mathematical Statistics* **1951**, *22* (1), 79–86.

Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* **1991**, *37* (1), 145–151.

Rényi, A. On Measures of Information and Entropy. *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960*. pp. 547–561, 1961.

Singhal, A. Modern Information Retrieval: A Brief Overview. Bulletin of the *IEEE Computer Society Technical Committee on Data Engineering* **2001**, *24* (4), 35–43.

INTENTIONALLY LEFT BLANK.

Appendix A. Divergence Measure Equations

The following details the divergence measure equations.

Let P be one text *corpus*, Q be another text *corpus*.

Let P_{types} = set of unique *types* in P, and Q_{types} = set of unique *types* in Q,

then let N be

$$| P_{types} \cup Q_{types} | = \#\text{unique } types \text{ in the union of sets P and Q} \quad \text{(A-1)}$$

For every unique *type* a(i), let p(i) be its *relative frequency* in P,

$$p(i) = (\#\text{occurrences of } token \text{ a(i) in P}) / (\text{total } \#tokens \text{ in P}) \quad \text{(A-2)}$$

and let q(i) be its *relative frequency* in Q

$$q(i) = (\#\text{occurrences of } token \text{ a(i) in Q}) / (\text{total } \#tokens \text{ in Q}) \quad \text{(A-3)}$$

Using these definitions, the equations for the divergence measures are as follows:

Rényi

$$(P;Q;\alpha) = \frac{1}{\alpha - 1} \log_2 \sum_{i=1}^{N} \left[p(i)^{1-\alpha} \cdot q(i)^{\alpha} \right] \quad \text{(A-4)}$$

Kullback-Leibler

$$(P;Q) = \sum_{i=1}^{N} \left[p(i) \log_2 \frac{p(i)}{q(i)} \right] \quad \text{(A-5)}$$

Bhattacharyya

$$(P;Q) = 1 - \sum_{i=1}^{N} \sqrt{p(i) \cdot q(i)} \quad \text{(A-6)}$$

Jensen-Shannon

$$(P;Q) = \frac{1}{2} \left[\sum_{i=1}^{N} p(i) \left(\log_2 p(i) - \log_2 \left(\frac{p(i) + q(i)}{2} \right) \right) + \sum_{i=1}^{N} q(i) \left(\log_2 q(i) - \log_2 \left(\frac{p(i) + q(i)}{2} \right) \right) \right] \quad \text{(A-7)}$$

Variational

$$(P;Q) = \sum_{i=1}^{N} \left| p(i) - q(i) \right| \quad \text{(A-8)}$$

Euclidean

$$(P;Q) = \sqrt{\sum_{i=1}^{N} \left(p(i) - q(i) \right)^2} \quad \text{(A-9)}$$

Cosine $$(P;Q) = 1 - \left(\frac{\sum_{i=1}^{N} p(i) \cdot q(i)}{\sqrt{\sum_{i=1}^{N} p(i)^2 \cdot \sum_{i=1}^{N} q(i)^2}} \right)$$ (A-10)

Appendix B. Punctuation

The following is a list of the punctuation recognized by the preprocessing option **Tokenize Punctuation**. Any punctuation not on this list will not be properly split off:

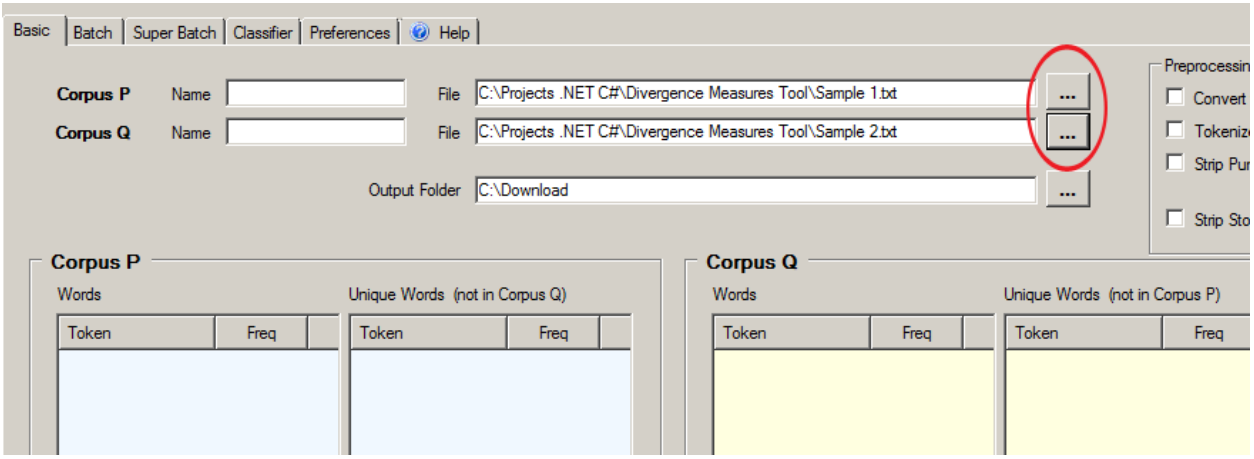
PUNCTUATION NAME	MARK	UNICODE
period	.	002E
comma	,	002C
question mark	?	003F
colon	:	003A
semi-colon	;	003B
exclamation mark	!	0021
left parentheses	(	0028
right parentheses	)	0029
forward slash	/	002F
back slash	\	005C
left bracket	[	005B
right bracket	]	005D
dash	-	002D
equal sign	=	003D
Arabic comma	،	060C
Arabic semi-colon	؛	061B
Arabic period	۔	061E
Arabic question mark	؟	061F
Left Single Quotation Mark	‘	2018
Single High-Reversed Quot Mark	‛	201B
Right Single Quotation Mark	’	2019
Left Double Quotation Mark	“	201C
Right Double Quotation Mark	”	201D
Double High-Reversed Quot Mark	‟	201F
Double Prime	″	2033
Reversed Prime		2035
Reversed Double Prime		2036

INTENTIONALLY LEFT BLANK.

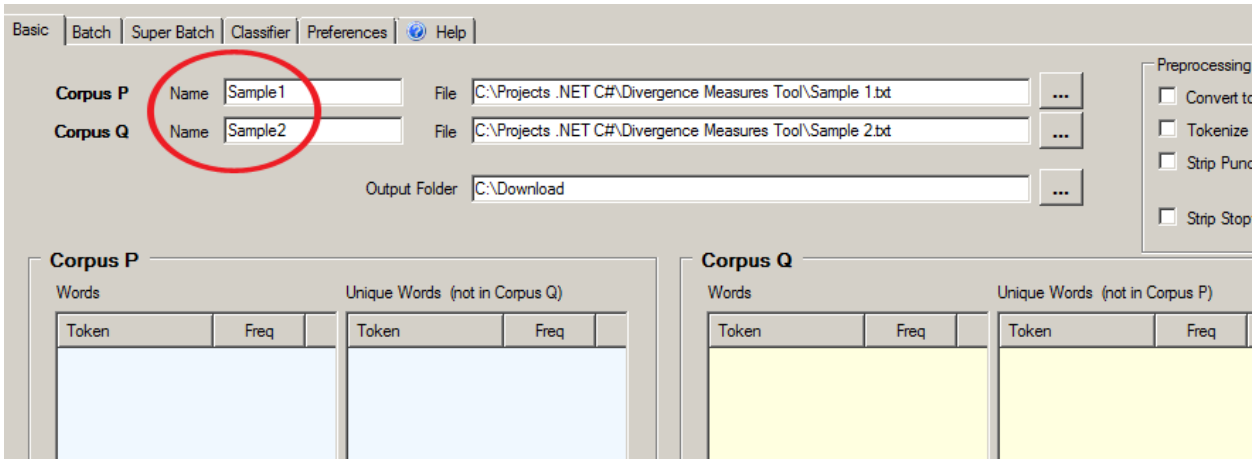
Appendix C. Tutorial

There are two example files included with this tool, named “Sample 1.txt” and “Sample 2.txt.” This tutorial will walk the user through using the **Basic** tab with these files, both so that the users can ensure the tool is working properly and so that users can understand how to run the tool with their own files.

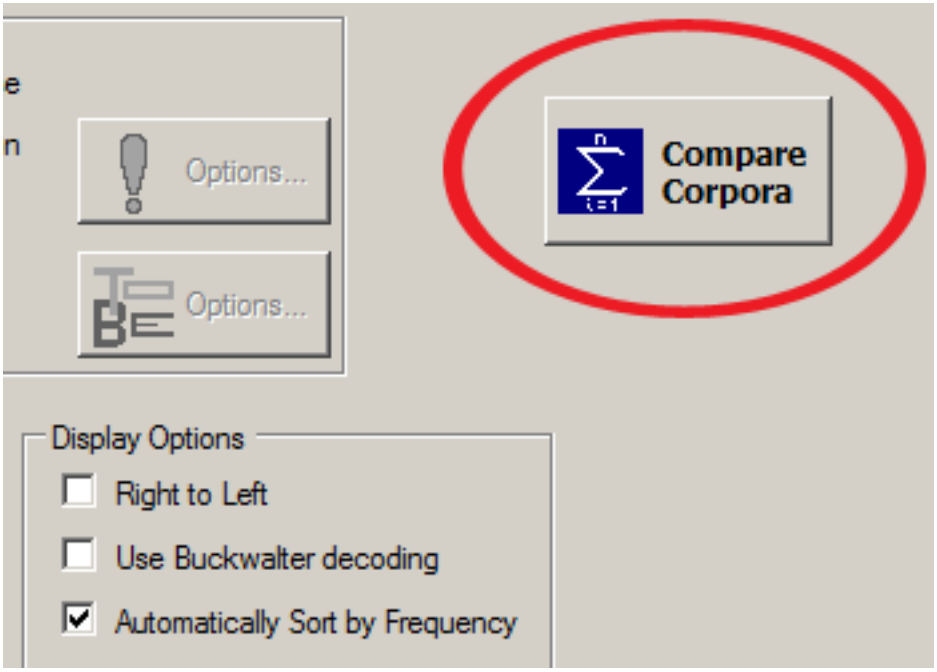
1. First, upload the files. Select the button labeled “…” to the right of **File** fields and navigate to the folder where the Divergence Measure Tool is stored, then select the example files.



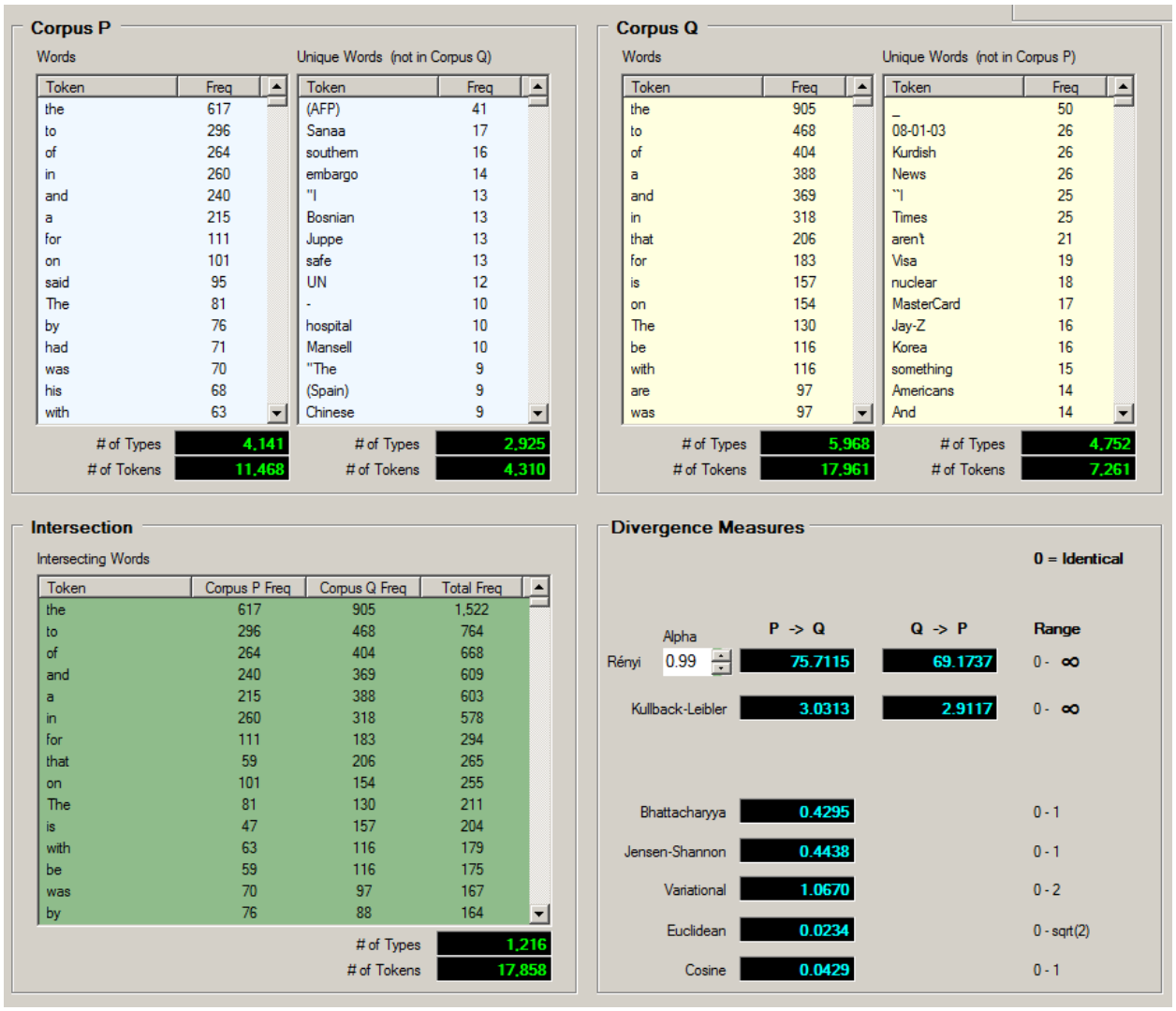
2. Now name the files by filling in the **Name** fields.



3. Click the **Compare Corpora** button.

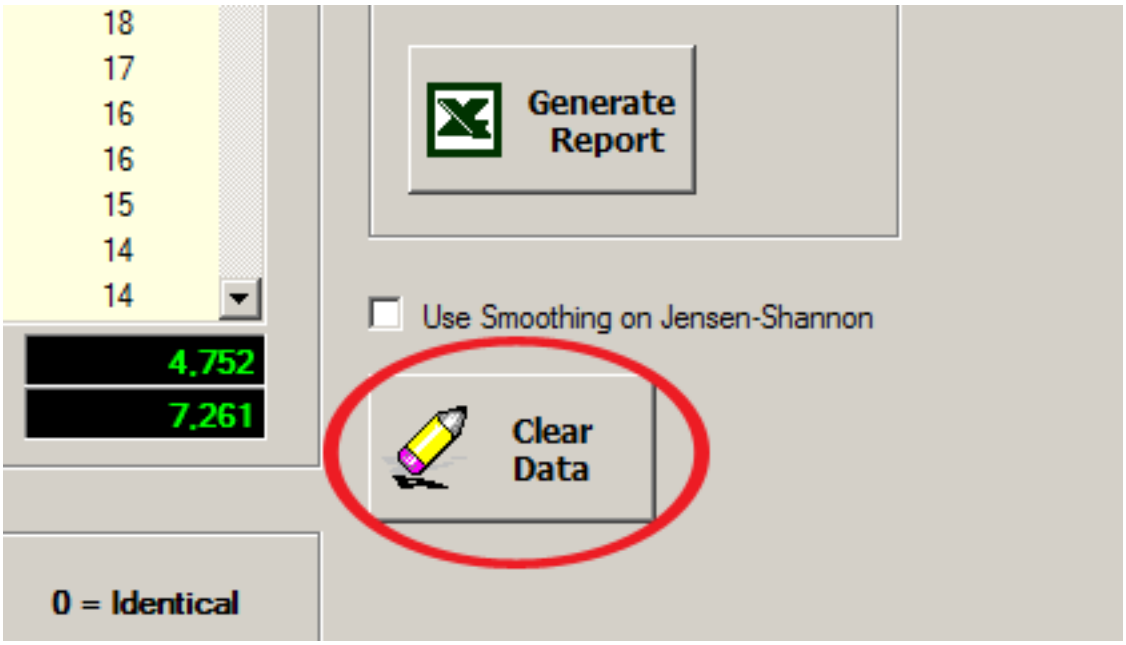


The results should look like this:

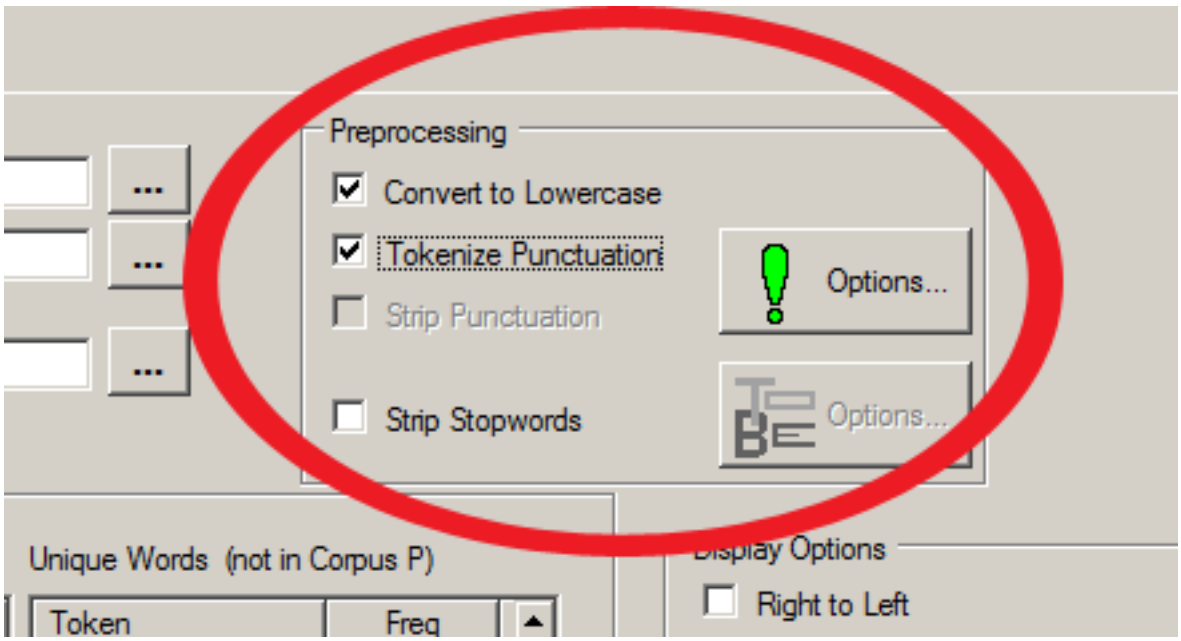


Notice anything strange? We didn't tokenize the punctuation or lowercase the text, so a single word (the) shows up as three different *types* – `the, The,` and `"The`.

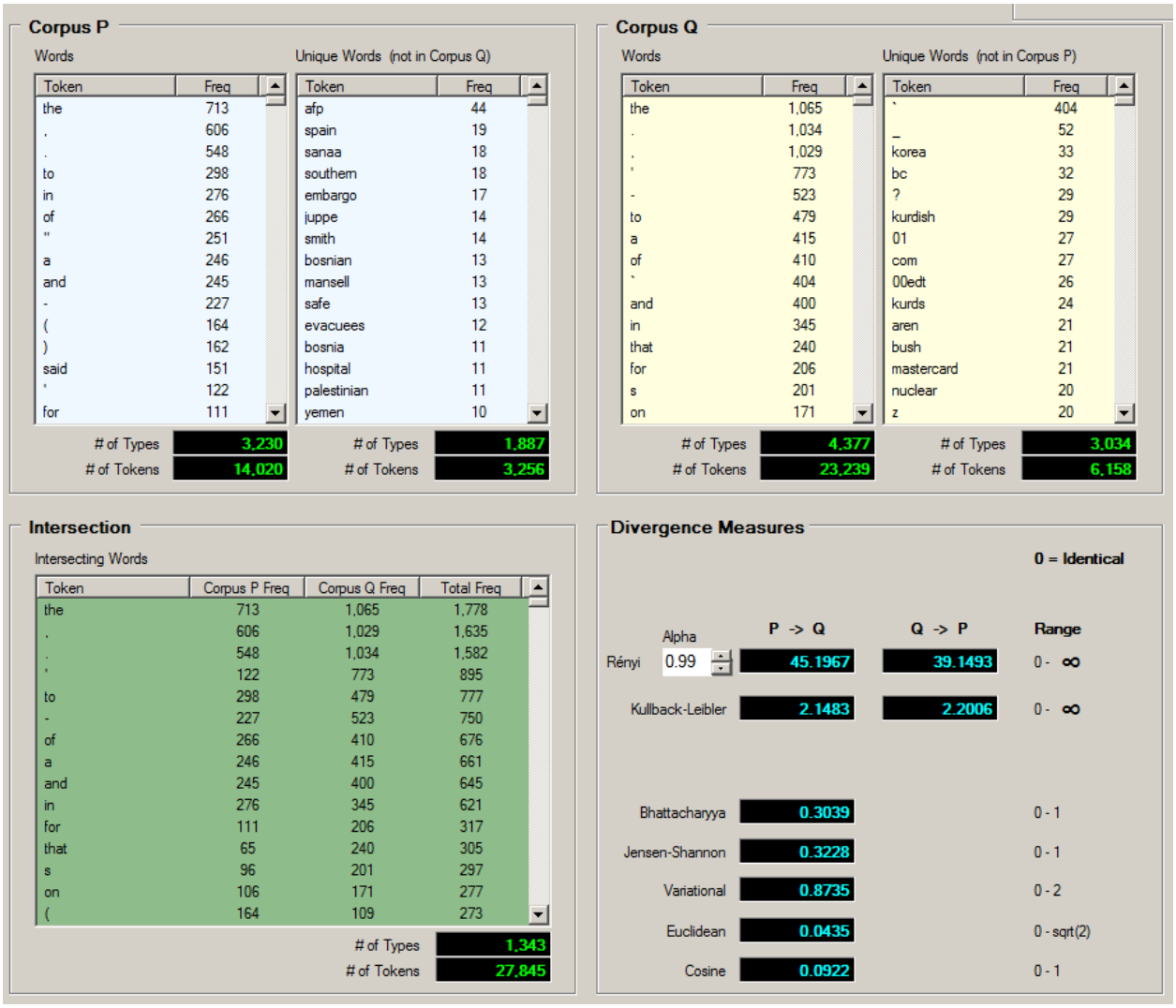
4. To fix this, click the **Clear Data** button, then repeat steps 1 and 2.



5. Now, select “Convert to Lowercase” and “Tokenize Punctuation” in the **Preprocessing** box.



6. Click the **Compare Corpora** button. Now the results should look like this:



Corpus P

Words

Token	Freq
the	713
,	606
.	548
to	298
in	276
of	266
"	251
a	246
and	245
-	227
(	164
)	162
said	151
'	122
for	111

# of Types: 3,230
# of Tokens: 14,020

Unique Words (not in Corpus Q)

Token	Freq
afp	44
spain	19
sanaa	18
southern	18
embargo	17
juppe	14
smith	14
bosnian	13
mansell	13
safe	13
evacuees	12
bosnia	11
hospital	11
palestinian	11
yemen	10

# of Types: 1,887
# of Tokens: 3,256

Corpus Q

Words

Token	Freq
the	1,065
.	1,034
,	1,029
"	773
-	523
to	479
a	415
of	410
`	404
and	400
in	345
that	240
for	206
s	201
on	171

# of Types: 4,377
# of Tokens: 23,239

Unique Words (not in Corpus P)

Token	Freq
`	404
_	52
korea	33
bc	32
?	29
kurdish	29
01	27
com	27
00edt	26
kurds	24
aren	21
bush	21
mastercard	21
nuclear	20
z	20

# of Types: 3,034
# of Tokens: 6,158

Intersection

Intersecting Words

Token	Corpus P Freq	Corpus Q Freq	Total Freq
the	713	1,065	1,778
,	606	1,029	1,635
.	548	1,034	1,582
'	122	773	895
to	298	479	777
-	227	523	750
of	266	410	676
a	246	415	661
and	245	400	645
in	276	345	621
for	111	206	317
that	65	240	305
s	96	201	297
on	106	171	277
(	164	109	273

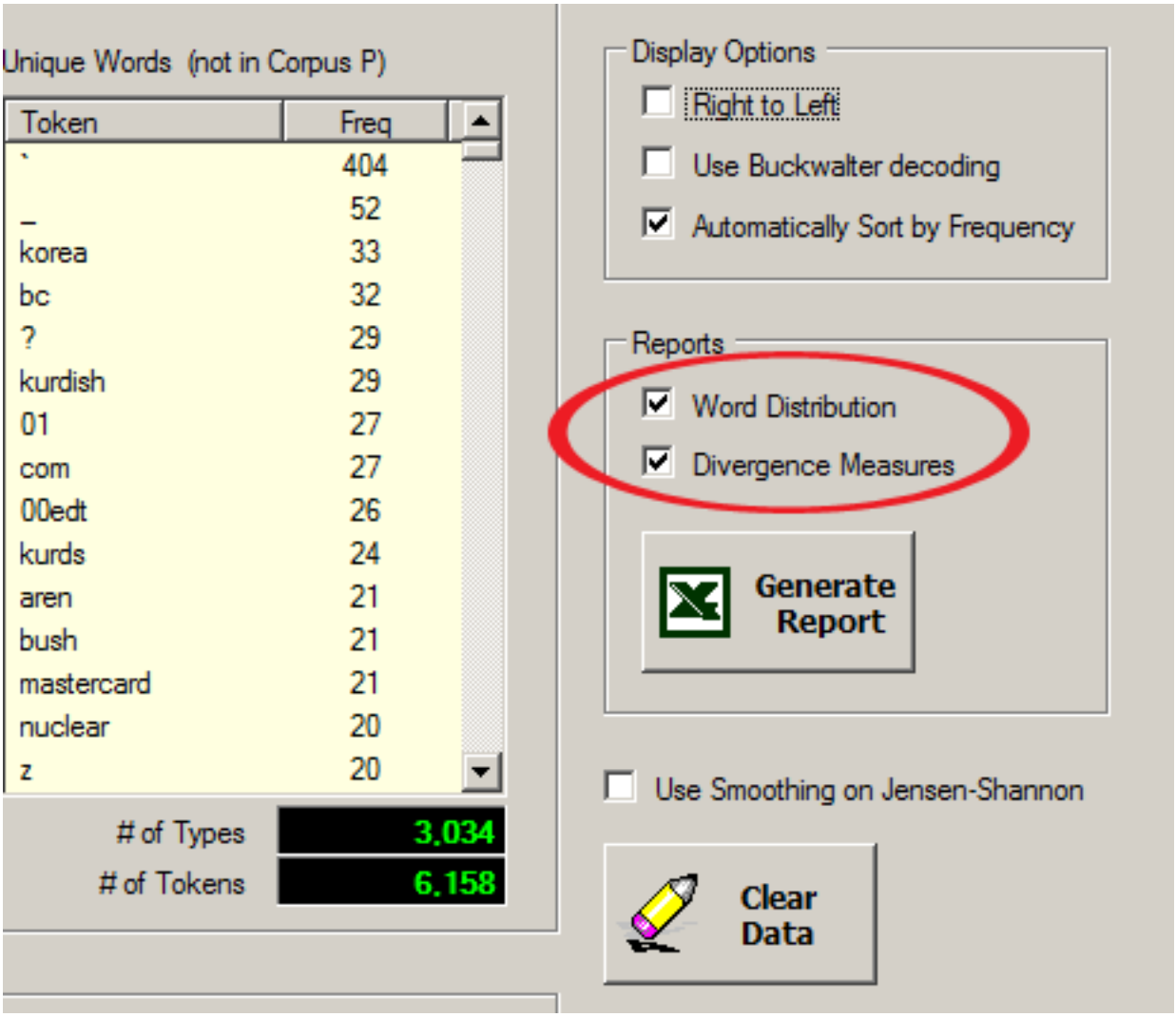
# of Types: 1,343
# of Tokens: 27,845

Divergence Measures

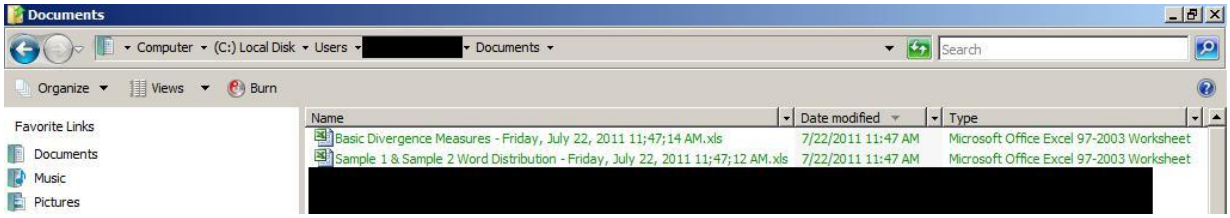
0 = Identical

	Alpha	P -> Q	Q -> P	Range
Rényi	0.99	45.1967	39.1493	0 - ∞
Kullback-Leibler		2.1483	2.2006	0 - ∞
Bhattacharyya		0.3039		0 - 1
Jensen-Shannon		0.3228		0 - 1
Variational		0.8735		0 - 2
Euclidean		0.0435		0 - sqrt(2)
Cosine		0.0922		0 - 1

7. Select both checkboxes in the **Results** box.



8. Click the **Generate Reports** button. This should create two spreadsheets in the Documents folder, named "Word Distribution" and "Divergence Measures."



Appendix D. FAQ

Q: Why are there different counts for the words `The` and `the`?

A: This tool is case-sensitive; if the user wants words with different casing to be counted together, select the **Convert to Lowercase** checkbox in the ***Preprocessing*** section prior to running the comparison.

Q: Why are there different counts for the words `the` and `"the`?

A: This tool looks for exact matches in order to recognize unique words. If the user wants to count the word and the punctuation preceding or following it separately, select the **Tokenize Punctuation** checkbox in the ***Preprocessing*** section prior to running the comparison.

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 GOVT PRNTG OFC
(PDF) A MALHOTRA

2 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CIO LL
IMAL HRA MAIL & RECORDS MGMT

5 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CII T
DOUGLAS M BRIESCH
CLAIRE E JAJA
TERRENCE J MOORE
CLARE R VOSS
BARBARA D BROOME